# *Travelling through time with corpus annotation software*

Paul Rayson
Computing Department
Lancaster University

# Not just my own work …

- Nicholas Smith
- Alistair Baron

- Dawn Archer

LANCASTER UNIVERSITY

uclan

# Motivation

Part 1

OPEN Content Alliance

INTERNET ARCHIVE

YAHOO!®

Microsoft®

Live Search

MBooks

EEBO EARLY ENGLISH BOOKS ONLINE

CHADWYCK-HEALEY

Google™ Book Search

UNIVERSITY OF OXFORD

DOMINUS ILLUMINATIO MEA

TEXT CREATION PARTNERSHIP

# OPEN Content Alliance

- Internet Archive
- December 2006
- 100,000 books on its servers
- Public access
- Opt in

- Microsoft Book Search
- Via MSN
- 100,000 books in the British Library
- Beta test in USA from December 2006

Microsoft

Live Search

- Small excerpts online due to copyright restriction … court case ongoing
- Opt out
- MBooks: entire collection at University of Michigan library
- Oxford University (1 million books of Bodleian Library)

- Digital facsimile page images of virtually every work printed in England, Ireland, Scotland, Wales from 1473-1700
- Separate initiative Text Creation Partnership is creating SGML versions for full text of 25,000 EEBO works

EEBO
EARLY ENGLISH BOOKS
ONLINE

CHADWYCK-HEALEY

TEXT CREATION
PARTNERSHIP

OPEN Content Alliance

INTERNET ARCHIVE

YAHOO!

Microsoft

Live Search

MBooks

Google Book Search

UNIVERSITY OF OXFORD

DOMINUS ILLUMINATIO MEA

EEBO
EARLY ENGLISH BOOKS ONLINE

CHADWYCK-HEALEY

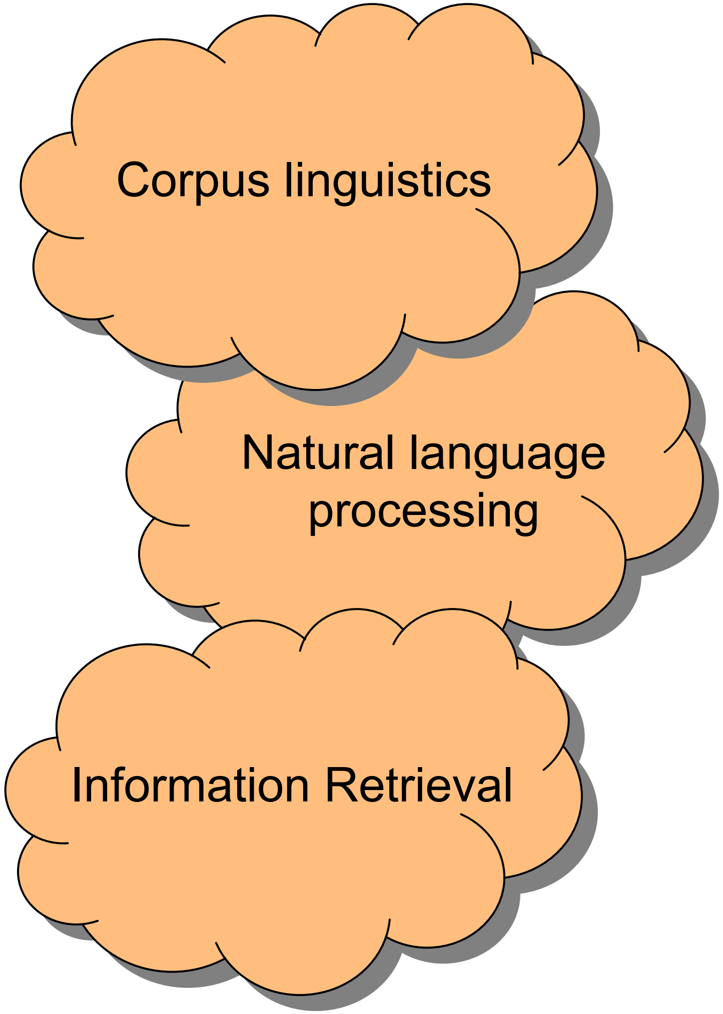
TEXT CREATION PARTNERSHIP

# Output from these initiatives

- Typically image based
- Some full text available for searching if not download
  - E.g. TCP data available to members
- Focussed on historical and out of copyright material

# Typical operations on modern data

- Annotation
  - POS tagging
- Retrieval
  - Frequency lists
  - N-grams
  - Search Engines
  - Concordances
  - Collocations


Corpus linguistics

Natural language processing

Information Retrieval

# We will need to carry out similar operations on historical data

- Historical corpus linguistics
- Search engines for new text collections and digital libraries
- Named entity extraction
- Historical text mining
- New research methods in History

# Problems faced when applying modern tools to historical data

Part 2

# Using automated systems of annotation on historical texts is problematic …

EModE texts pose the following “problems”:

- Archaic *–eth* and *–(e)st* verb suffixes, e.g. *doth*, *hath*, *hast*, *sayeth*, etc., which persist in specialised contexts: religious and poetic usage
- Fused forms, e.g. *’Tis* (*It is*)
- Spellings that are variable even in modern-day usage, e.g. *center/centre*, skilful/skillful/skilfull, the suffixes *-or*/*-our*, *-ise*/*-ize*
- Archaic forms like *howbeit*, *betwixt*, for which no obvious modern equivalent exists
- Compound words, e.g. *it self*, *now adays*, *in stead*
- Proper names of Latin origin that are sometimes modernised, e.g. *Galilaeo* (*Galileo*)
- In consequence ... the results generated by existing software are not always robust!

# Accuracy and robustness

- POS taggers tested across registers and genres of modern data for coverage and accuracy
- Less is known about their accuracy on historical data
- Spelling issues
    - Modern: hyphenation and tokenisation
    - Historical: different conventions, compositing practices

# Possible solutions

Part 3

# Previous work in …

Information Retrieval

- Fuzzy search engine
- Aimed at successful retrieval for novice users without expertise in the text
- Expand the query term using known letter replacements
- Text can't be pre-indexed
- 100% recall important, precision obtained via sorting results

# Previous work in …

Corpus linguistics

Natural language processing

- Adding historical variants to POS tagger’s lexicon
  - E.g. TreeTagger application to GerManC
- Back-dating lexicon
  - E.g. ENGCG application to Helsinki corpus

# Our scenario

- Apply to a number of techniques
- POS and Semantic tagging
- Frequency profiling, n-grams etc

- Crucially – most previous approaches don’t deal with contextual variants

# Our response

…to further develop an existing Modern Tagger (= the UCREL Semantic Annotation System)

… USAS automatically annotates present-day texts (spoken and written) …

# Semantic fields captured by the tagger(s)

**Hierarchy of 21 major discourse fields (see below), which expands into 232 semantic field tags:**

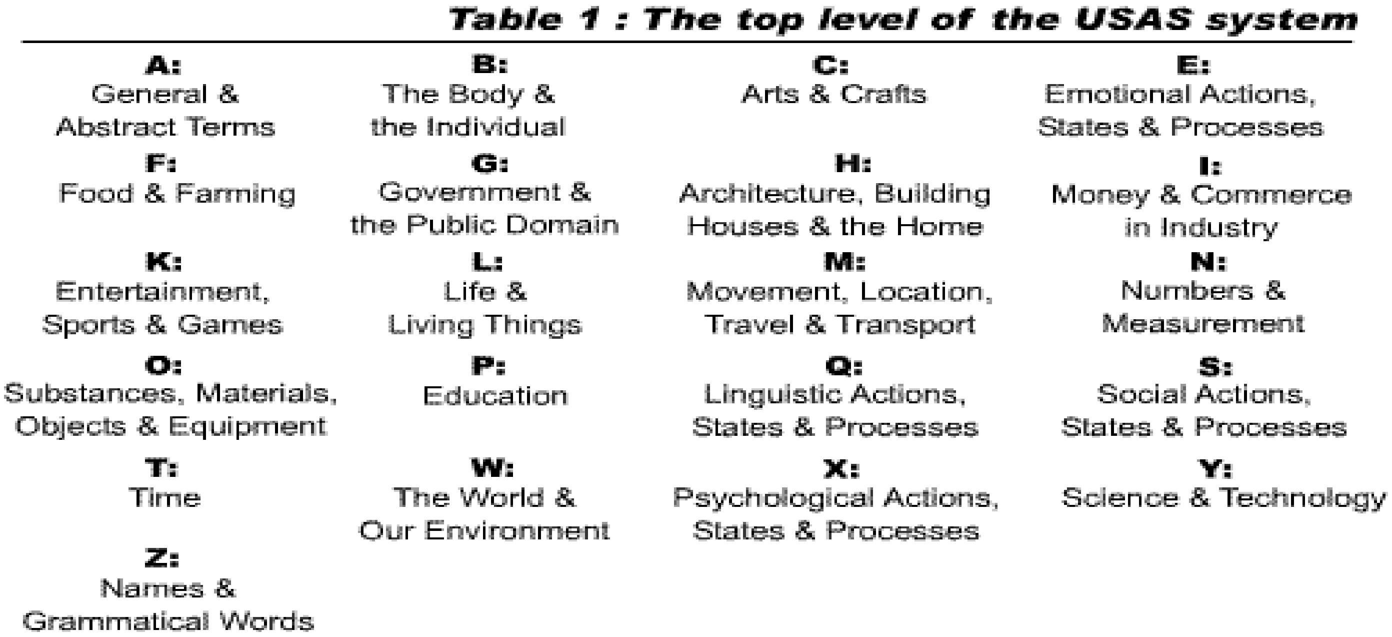


*Table 1 : The top level of the USAS system*

| | | | |
|---|---|---|---|
| **A:**<br>General & Abstract Terms | **B:**<br>The Body & the Individual | **C:**<br>Arts & Crafts | **E:**<br>Emotional Actions, States & Processes |
| **F:**<br>Food & Farming | **G:**<br>Government & the Public Domain | **H:**<br>Architecture, Building Houses & the Home | **I:**<br>Money & Commerce in Industry |
| **K:**<br>Entertainment, Sports & Games | **L:**<br>Life & Living Things | **M:**<br>Movement, Location, Travel & Transport | **N:**<br>Numbers & Measurement |
| **O:**<br>Substances, Materials, Objects & Equipment | **P:**<br>Education | **Q:**<br>Linguistic Actions, States & Processes | **S:**<br>Social Actions, States & Processes |
| **T:**<br>Time | **W:**<br>The World & Our Environment | **X:**<br>Psychological Actions, States & Processes | **Y:**<br>Science & Technology |
| **Z:**<br>Names & Grammatical Words | | | |

**Presently exploring ways in which we may need to alter/ amend the 232 categories for the Historical Semantic Tagger**

**– this work will also draw on Shakespearean Thesauri (i.e. Spevack 1993, Trussler 1986) for Early Modern period**

# The Structure of the Modern Tagger

Incorporates "modern" lexical resources, i.e. a list of single word forms and multi-word units (MWUs)

… which are fed into a PART-OF-SPEECH and SEMANTIC tagger …

Part-of-speech tags are assigned to <u>every</u> lexical item or multi-word expression (MWE), using probabilistic Markov models of likely part-of-speech sequences (- 97% accuracy)

POS TAGGER

CONTEMPORARY LEXICON

CONTEMPORARY MWE LIST

SEM TAGGER

The output is fed into SEMTAG, which assigns tags on the basis of pattern matching between the text and the two computer dictionaries (- 92% accuracy)

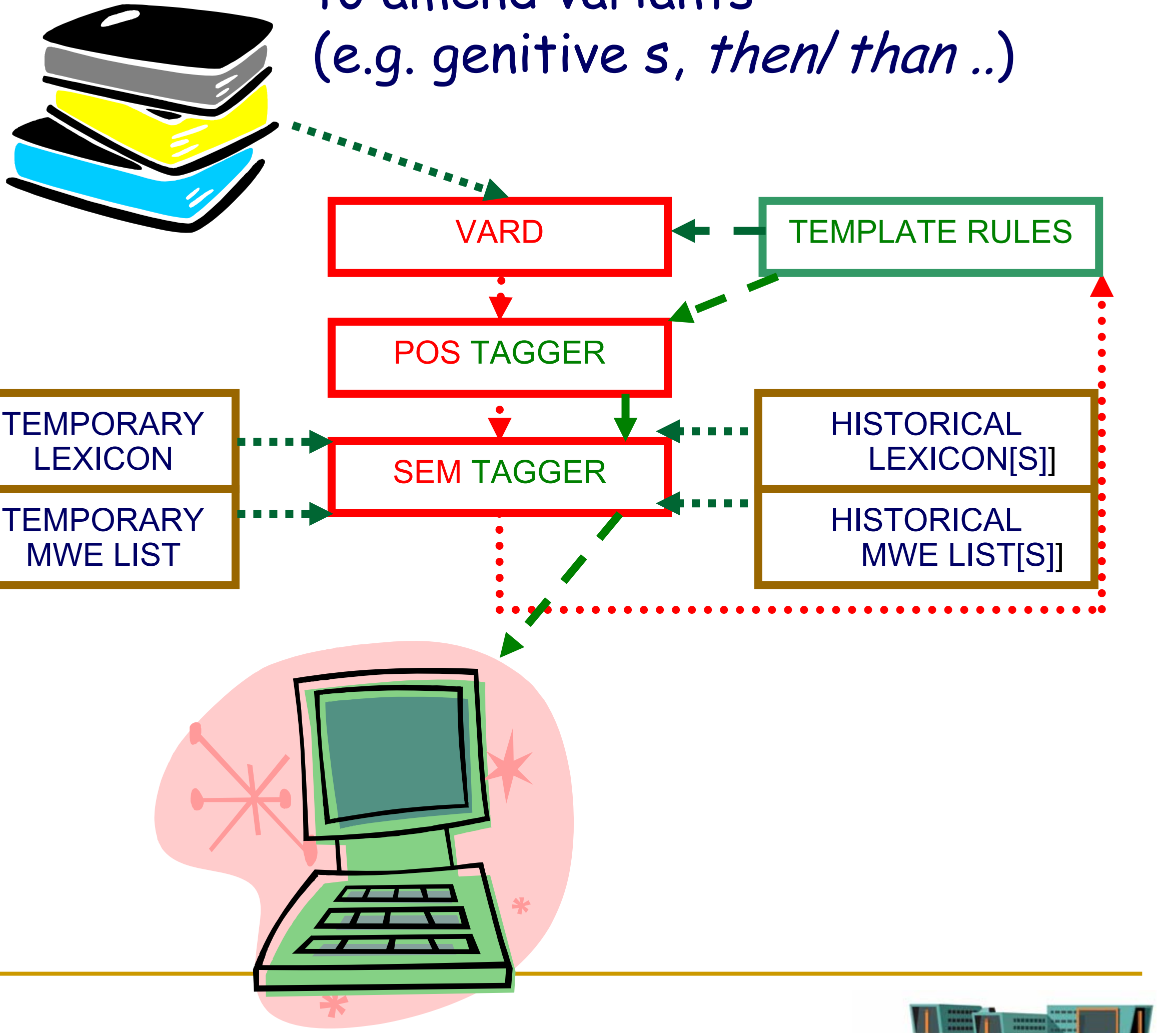
# The Structure of the Historical Tagger

Incorporates:

Additional lexicons, separated according to period (16-17 C, 18-19 C, 20-21 C)

... a VARiant Detector (= a spelling detector and normaliser)

... and a component that allows us to use the context to amend variants (e.g. genitive s, *then/ than* ..)



# An important point about the VARD

Although the VARD allows for the detection and "normalisation" of variants to their modern equivalents, it should be noted that ...

- The original variants are retained in the text
- We're not carrying out spell checking per se (no "correct" spelling in EmodE period) ...
  - Rather, our ultimate aim is to develop a system that does not merely offer the user possible "suggestions" for spelling variants (as in the case of MS-Word and Aspell), but *automatically* regularises variants within a text to their modernised forms so that historical corpora become more amenable to further annotation and analysis.

# VARD uses a hybrid approach to match EmodE variants to modern equivalents

- Version 1
  - Known variants list

- Version 2
  - Soundex
  - Edit distance
  - Letter replacement heuristics

- Version 3
  - Contextual rules

# Known variants list

= A search and replace script and a list of terms, which "matches" spelling variants to their "normalised" equivalents:

- Presently contains 45,805 entries
- With several categories: "o", "m", "mod", "d", "f", etc.
- Manually constructed (although labour intensive, has proved to be accurate: see Rayson et al., 2005)

# Soundex match (O’Dell and Russell 1918)

… Identifies strings that sound similar regardless of their spelling …

1. Replace all but the first letter with the digit listed below:

| | |
|---|---|
| 0: | A, E, I, O, U, H, W, Y |
| 1: | B, F, P, V |
| 2: | C, G, J, K, Q, S, X, Z |
| 3: | D, T |
| 4: | L |
| 5: | M, N |
| 6: | R |

2. Remove any pairs of digits that are the same and occur next to each other in the string.
3. Remove all occurrences of the digit 0.
4. The Soundex code is the first 4 letters of the remaining string.

’disapont’ and ‘disappoint’ both have code D215

But so do ‘dispense’, ‘deceiving’ and ‘despond’

# Edit distance

- Levenshtein distance (1965)

= Measure of similarity between two strings

- ‘disapont’ -> ‘disap**p**o**i**nt’ **distance = 2:**  
  insertion: p  
  insertion: i

- ‘dis**a**p**o**n**t**’ -> ‘disp**e**n**se**’ **distance = 4:**  
  deletion: a  
  substitution: o → e  
  substitution: t → s  
  insertion: e

- ‘d**isapo**n**t**’ -> ‘d**eceiv**in**g**’ **distance = 7:**  
  substitution: i → e  
  substitution: s → c  
  substitution: a → e  
  insertion: i  
  substitution: p → v  
  substitution: o → i  
  substitution: t → g

# Letter replacements

- Manually constructed - based on corpus data
- 51 rules, some specifying ‘context’ for replacement
  - Replace final ck with c
  - Replace u with v
  - Replace v with u
  - Replace final ‘d with ed
  - Remove final e

# Starting to automatically derive these

- From 45K known variant (types)
  - Edit distance 1: 27067
  - Edit distance 2: 11918
  - Edit distance 3: 4350
  - Edit distance 4: 897
  - Edit distance 5+: 216
- Frequencies of letter replacements
  - e >> _: 6501
  - ‘ >> e: 2730
  - y >> i: 2602
  - u >> v: 1662
  - …

# Contextual rules

- A component to cope with inconsistencies (orthographical and other) that can only be disambiguated via the “context”

- Contextual rules
  - then/than, bee/be, doe/do
  - Apostrophes

- Uses context rules, such as ‘if … then’, e.g. …

If the input consists of:

| | |
|---|---|
| her | tagged as APPGE (possesive pronoun) |
| Majesties | tagged as NN2 (plural noun) |

Then: change the word

| | |
|---|---|
| Majesties to ... | Majesty's (sing. noun+genitive) |

**NOTE:- we also intend to make use of *semantic* info.**

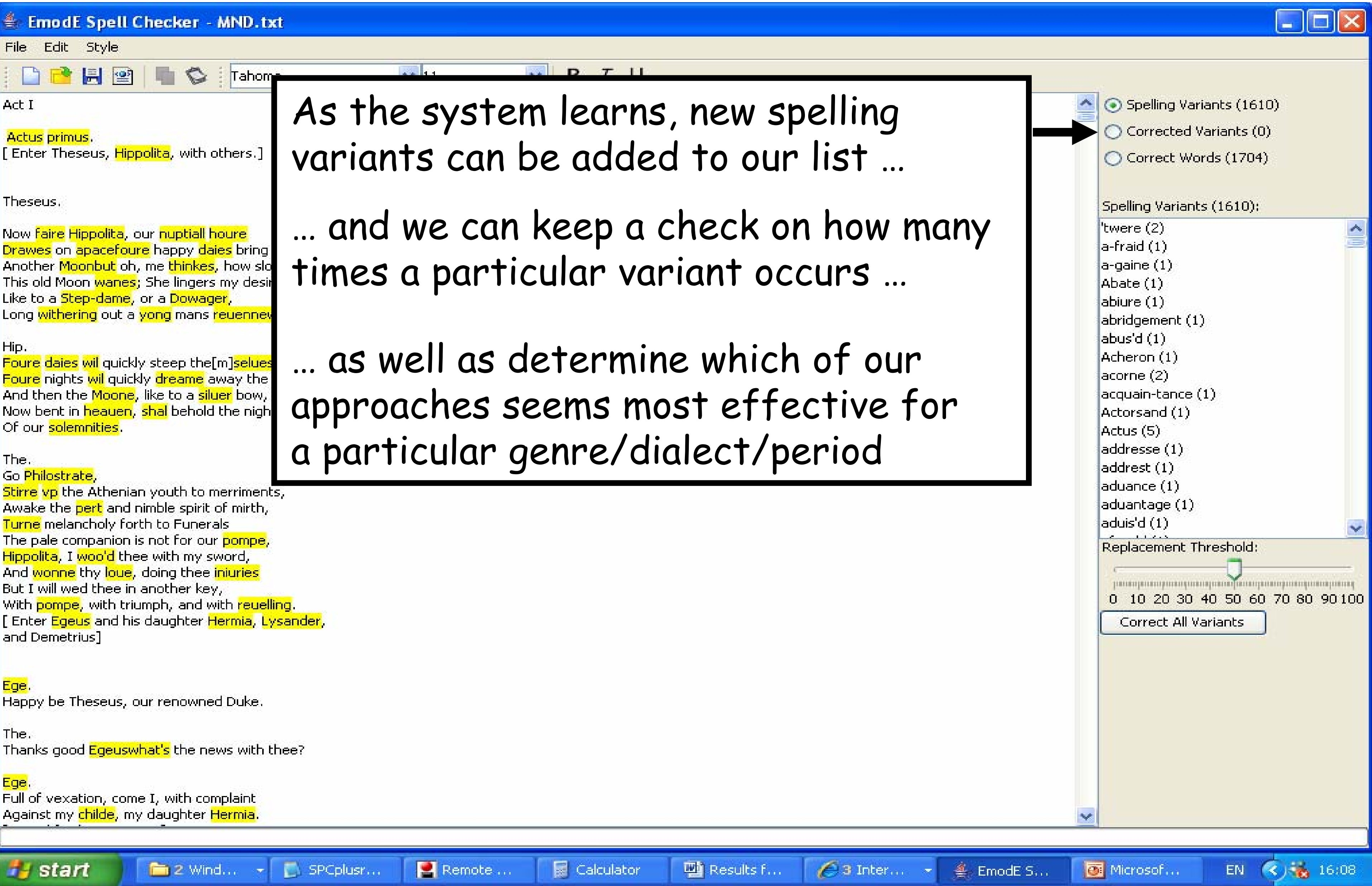
# Machine learning

- Trained by manual additions to the dictionary

- Weighting of different approaches changes during the use of the system …

- e.g. when applied to Shetland component of SCOTS corpus, Soundex is preferred over known variants

# Training the system to learn as it normalises …



Act I

Actus primus.
[ Enter Theseus, Hippolita, with others.]

Theseus.

Now faire Hippolita, our nuptiall houre
Drawes on apacefoure happy daies bring
Another Moonbut oh, me thinkes, how slo
This old Moon wanes; She lingers my desi
Like to a Step-dame, or a Dowager,
Long withering out a yong mans reuennew

Hip.
Foure daies wil quickly steep the[m]selues
Foure nights wil quickly dreame away the
And then the Moone, like to a siluer bow,
Now bent in heauen, shal behold the nigh
Of our solemnities.

The.
Go Philostrate,
Stirre vp the Athenian youth to merriments,
Awake the pert and nimble spirit of mirth,
Turne melancholy forth to Funerals
The pale companion is not for our pompe,
Hippolita, I woo'd thee with my sword,
And wonne thy loue, doing thee iniuries
But I will wed thee in another key,
With pompe, with triumph, and with reuelling.
[ Enter Egeus and his daughter Hermia, Lysander,
and Demetrius]

Ege.
Happy be Theseus, our renowned Duke.

The.
Thanks good Egeuswhat's the news with thee?

Ege.
Full of vexation, come I, with complaint
Against my childe, my daughter Hermia.

● Spelling Variants (1610)
○ Corrected Variants (0)
○ Correct Words (1704)

Spelling Variants (1610):

'twere (2)
a-fraid (1)
a-gaine (1)
Abate (1)
abiure (1)
abridgement (1)
abus'd (1)
Acheron (1)
acorne (2)
acquain-tance (1)
Actorsand (1)
Actus (5)
addresse (1)
addrest (1)
aduance (1)
aduantage (1)
aduis'd (1)

Replacement Threshold:

0 10 20 30 40 50 60 70 80 90 100

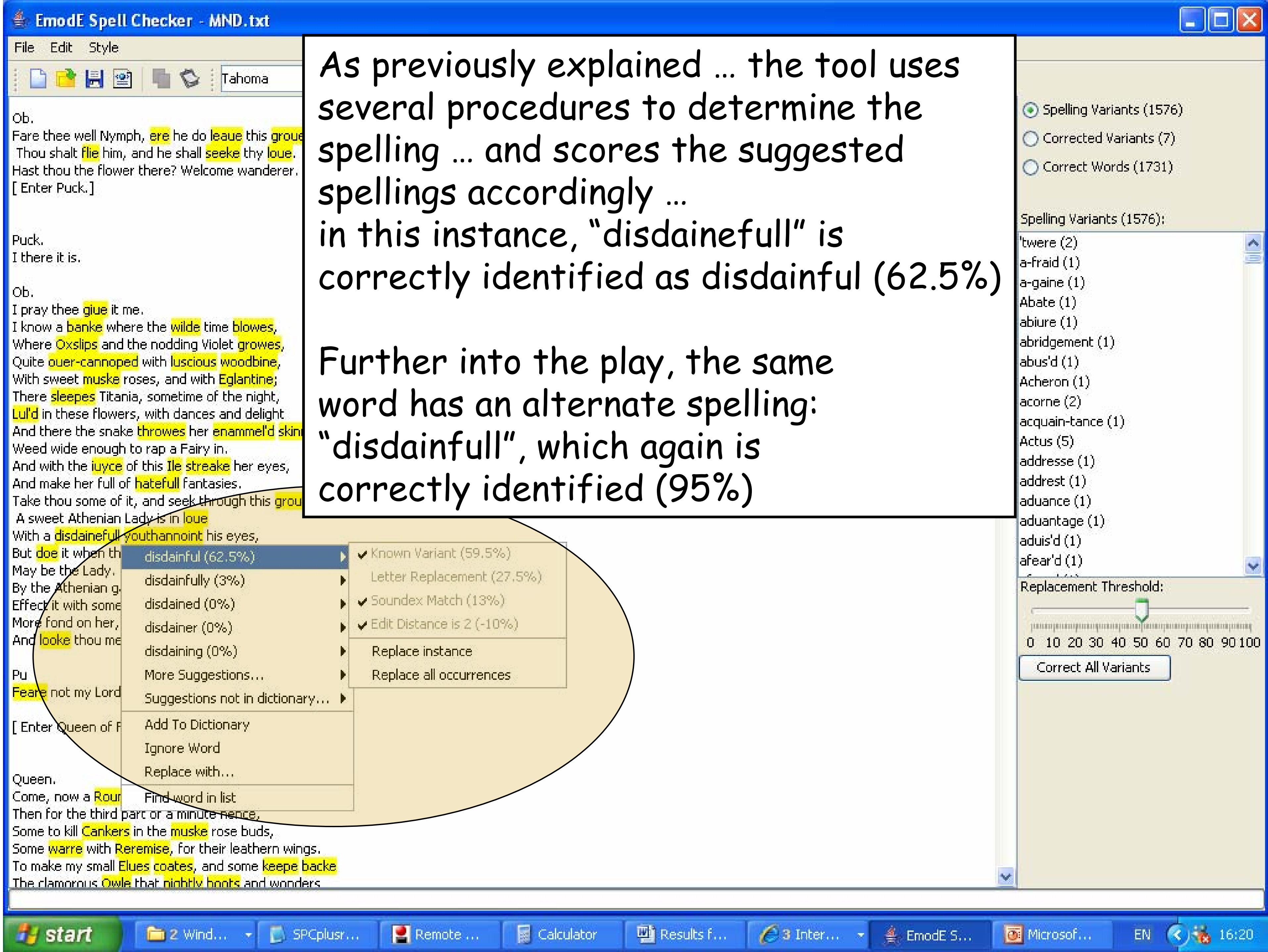
As the system learns, new spelling variants can be added to our list …

… and we can keep a check on how many times a particular variant occurs …

… as well as determine which of our approaches seems most effective for a particular genre/dialect/period

As previously explained … the tool uses several procedures to determine the spelling … and scores the suggested spellings accordingly …
in this instance, “disdainefull” is correctly identified as disdainful (62.5%)

Further into the play, the same word has an alternate spelling: “disdainfull”, which again is correctly identified (95%)



● Spelling Variants (1576)
○ Corrected Variants (7)
○ Correct Words (1731)

Ob.
Fare thee well Nymph, ere he do leaue this groue
Thou shalt flie him, and he shall seeke thy loue.
Hast thou the flower there? Welcome wanderer.
[ Enter Puck.]

Puck.
I there it is.

Ob.
I pray thee giue it me.
I know a banke where the wilde time blowes,
Where Oxslips and the nodding Violet growes,
Quite ouer-cannoped with luscious woodbine,
With sweet muske roses, and with Eglantine;
There sleepes Titania, sometime of the night,
Lul'd in these flowers, with dances and delight
And there the snake throwes her enammel'd skin
Weed wide enough to rap a Fairy in.
And with the iuyce of this Ile streake her eyes,
And make her full of hatefull fantasies.
Take thou some of it, and seek through this grou
A sweet Athenian Lady is in loue
With a disdainefull youthannoint his eyes,
But doe it when th
May be the Lady.
By the Athenian g
Effect it with some
More fond on her,
And looke thou me

Pu
Feare not my Lord

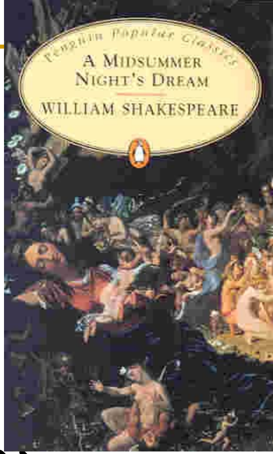
[ Enter Queen of F

Queen.
Come, now a Roun
Then for the third part of a minute hence,
Some to kill Cankers in the muske rose buds,
Some warre with Reremise, for their leathern wings.
To make my small Elues coates, and some keepe backe
The clamorous Owle that nightly hoots and wonders.

# Evaluation

Part 4

# Some preliminary results …



No. of variants initially found in MND by VARD = 1610. A quick check of the variants revealed that a handful of these were "real" words that VARD had not recognised (because of not being in our list (=BNC Written Sampler))

Some real words were LATINATE terms … our present approach is to ignore these.

Others were NAMES of CHARACTERS … we tend to add these to the existing list.

The majority of "real" words were words still in use today, but which are not found in the BNC Written Sampler … consequently, we are interested in incorporating a more comprehensive word list …

# First 150 variants

VARD was able to offer appropriate suggestions for 149.
The first suggestion tended to be the right one …

.. with the exception of "vnhardned" … a possible solution here is to affix-strip.

## Types of variant "normalised" (from 150 list):

| | |
|---|---|
| u – v | e.g. aduis'd (1), beleeue (5), haue (95), leaue (15) |
| v – u | e.g. vrg'd (1), vs (21), vsuall (1), voyce (5), vp (26) |
| ie-y | e.g. chastitie (1), daies (3) |
| i – j | e.g. iewels (1), iniuries (1), iudgment (1) |
| Extra e | e.g. asleepe (5), Bottome (14), confesse (3) |
| 'd | e.g. chang'd (2), adus'd (1), bewitch'd (1) |
| Double ll | e.g. beautifull (1) |

Also normalised *apricocks* to *apricots*, *acquain-tance* to *acquaintance*, etc.

# Variation that VARD deals with successfully …

Apostrophes signalling missing letter(s) or sound(s): *'fore* ("before"), *hee'l* ("he will"),

Irregular apostrophe usage: *again'st* ("against"), *whil'st* ("whilst")

Contracted forms: *'tis* ("it is"), *thats* ("that is"), *youle* ("you will"), *t'anticipate* (" to anticipate")

Hyphenated forms: *acquain-tance* ("acquaintance")

Variation due to different use of graphs: <v>, <u>, <i>, <y>

Doubling of vowels and consonants – e.g. <-oo-> <-ll>

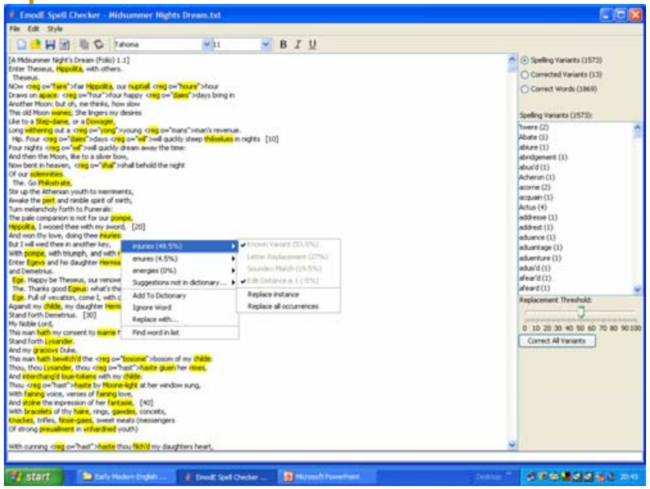
## Phenomena that is proving more problematic:

*I* to represent *aye* (= "yes")
Contraction of "stand-alone" words (e.g. *shalbe*)
Compounds that are now open (e.g. *Townes-men*)
Compounds that were then open (e.g. *our selues*)
Capitalisation (but useful as a "noun" marker?)

# Where next with the prototype …?

- The prototype is not yet making use of the contextual rules we've developed to cope with inconsistencies relating to the genitive and "then" versus "than", etc.
- These contextual rules rely on part-of-speech information
- Derive new letter replacement rules from training corpus and known variants list
- In addition …
  - We want to make use of semantic domain information as a means of disambiguating which variant forms belong to which normalised forms in instances where a one-to-one mapping isn't feasible – e.g. *piece/peace* and *peece*
  - We are considering whether the inclusion of etymological information might provide a further means of choosing between possible variants – by, for example, helping us to eliminate some variant-to-head word mappings if they cannot occur in a particular century …?

# The user's experience …



[A Midsummer Night's Dream (Folio) 1.1]
Enter Theseus, Hippolita, with others.
Theseus.
NOw <reg o="faire">fair Hippolita, our nuptiall <reg o="houre">hour
Draws on apace: <reg o="four">four happy <reg o="daies">days bring in
Another Moon: but oh, me thinks, how slow
This old Moon wanes; She lingers my desires
Like to a Step-dame, or a Dowager,
Long withering out a <reg o="yong">young <reg o="mans">man's revenue.
Hip. Four <reg o="daies">days <reg o="wil">will quickly steep themselues in nights [10]
Four nights <reg o="wil">will quickly dream away the time:
And then the Moon, like to a silver bow,
Now bent in heaven, <reg o="shal">shall behold the night
Of our solemnities.
The. Go Philostrate,
Stir up the Athenian youth to merriments,
Awake the pert and nimble spirit of mirth,
Turn melancholy forth to Funerals:
The pale companion is not for our pompe,
Hippolita, I wooed thee with my sword, [20]
And won thy love, doing thee iniuries
But I will wed thee in another key,
With pompe, with triumph, and with r
Enter Egeus and his daughter Hermia
and Demetrius.
Ege. Happy be Theseus, our renowr
The. Thanks good Egeus: what's the
Ege. Full of vexation, come I, with c
Against my childe, my daughter Herm
Stand forth Demetrius. [30]
My Noble Lord,
This man hath my consent to marrie h
Stand forth Lysander.
And my gracious Duke,
This man hath bewitch'd the <reg o="bosome">bosom of my childe:
Thou, thou Lysander, thou <reg o="hast">haste giuen her rimes,
And interchang'd loue-tokens with my childe:
Thou <reg o="hast">haste by Moone-light at her window sung,
With faining voice, verses of faining love,
And stolne the impression of her fantasie, [40]
With bracelets of thy haire, rings, gawdes, conceits,
Knackes, trifles, Nose-gaies, sweet meats (messengers
Of strong preuailment in vnhardned youth)

With cunning <reg o="hast">haste thou filch'd my daughters heart,

● Spelling Variants (1573)
○ Corrected Variants (13)
○ Correct Words (1869)

Spelling Variants (1573):

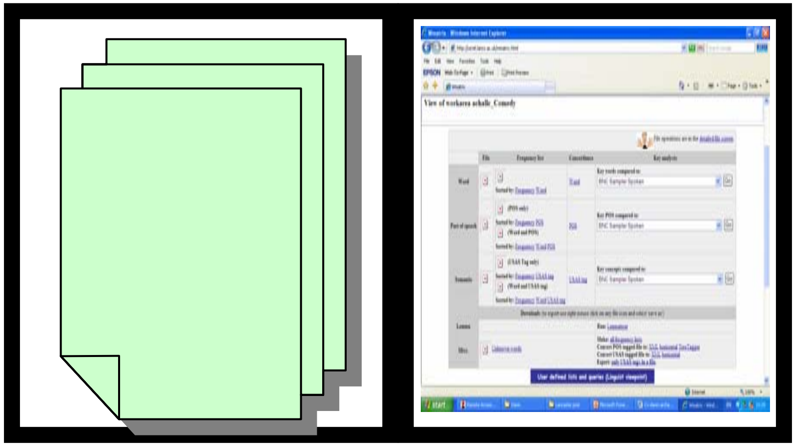
'twere (2)
Abate (1)
abiure (1)
abridgement (1)
abus'd (1)
Acheron (1)
acorne (2)
acquain (1)
Actus (4)
addresse (1)
addrest (1)
aduance (1)
aduantage (1)
aduenture (1)
adus'd (1)
afear'd (1)
afeard (1)

Replacement Threshold:

The user will utilise the VARD to detect and normalise spelling variants … at which point, the user will be given the option of part-of-speech tagging and semantically tagging their chosen text(s)

Once the text has been tagged, the user will have access to a split screen interface …

One window will provide an option to view the text (*in its original state or in its amended state*)



The remaining window will allow users to perform a number of searches … at the word, P-O-S and semantic level

# Summary

Part the last

# Summary and research potential

- When you go travelling in time with corpus software …

  Matching variant spellings (and other variant forms) to their “normalised” equivalent(s) means more meaningful results for those who want to analyse their datasets using standard corpus linguistic techniques (frequency profiles, concordances, collocations, extraction of n-grams, tagging)

- We know that we have to deal with altering the taxonomies used at POS and Semantic level to reflect changes in grammar and meaning over time

# Future possibilities ...?

- The VARD also allows for the exploration of spelling variation systematically. This might be across different centuries and/or across different text-types
- We would like to explore the feasibility of adapting the VARD so that it can "normalise":
  - Historical periods that are pre-Shakespeare
  - Dialectal variation in Pres-Day texts

# Thank you for your interest !

Contact details: Paul Rayson (paul@comp.lancs.ac.uk)

Further details re VARD and the Historical Tagger, available at: http://www.comp.lancs.ac.uk/ucrel/

Work presented here was carried out within the following project:

- *Scragg Revisited* funded by the British Academy (under the small research grant scheme)